

Efficient fluctuation analysis of biochemical pathways beyond the linear noise approximation using iNA

Philipp Thomas^{*†‡¶}, Hannes Matuschek^{§¶} and Ramon Grima^{*†}

^{*}*School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom*

[†]*SynthSys Edinburgh, University of Edinburgh, Edinburgh, United Kingdom*

[‡]*Department of Physics, Humboldt University of Berlin, Berlin, Germany*

[§]*Institute of Physics and Astronomy, University of Potsdam, Potsdam, Germany*

[¶]*These authors contributed equally.*

Abstract—The linear noise approximation is commonly used to obtain intrinsic noise statistics such as Fano factors and coefficients of variation for biochemical networks. These estimates are accurate for networks with large numbers of molecules. However it is well known that many biochemical networks are characterized by at least one species with a small number of molecules. We here describe modifications to the software intrinsic Noise Analyzer (iNA) which enable it to accurately compute noise statistics over wide ranges of molecule numbers. This is achieved by calculating the next order corrections to the linear noise approximation's estimates of variance and covariance of concentration fluctuations. The efficiency of the methods is significantly improved by automated just-in-time compilation using the LLVM framework leading to a fluctuation analysis which typically outperforms that obtained by means of exact stochastic simulations. iNA is hence particularly well suited for the needs of the computational biology community.

Keywords—Stochastic modeling; Linear Noise Approximation; genetic regulatory circuits

I. INTRODUCTION

Experimental studies have shown that the protein abundance varies from few tens to several thousands per protein species per cell [1]. It is also known that the standard deviation of the concentration fluctuations due to the random timing of molecular events (intrinsic noise) roughly scales as the square root of the mean number of molecules [2]. Hence it is expected that intrinsic noise plays an important role in the dynamics of those biochemical networks characterized by at least one species with low molecule numbers.

The stochastic simulation algorithm (SSA) is the conventional means of probing stochasticity in biochemical reaction systems [3]. This method simulates every reaction event and hence is typically slow for large reaction networks; this is particularly true if one is interested in intrinsic noise statistics which require considerable ensemble averaging of the trajectories produced by the SSA. A different route of inferring the required statistics involves finding an approximate solution of the chemical master equation (CME), a set of differential equations for the probabilities of the states of the system, which is mathematically equivalent to the SSA.

We recently developed intrinsic Noise Analyzer (iNA) [4], the first software package enabling a fluctuation analysis of biochemical networks via the Linear Noise Approximation (LNA) and Effective Mesoscopic Rate Equation (EMRE) approximations of the CME. The former gives the variance and covariance of concentration fluctuations in the limit of large molecule numbers while the latter gives the mean concentrations for intermediate to large molecule numbers and is more accurate than the conventional Rate Equations (REs).

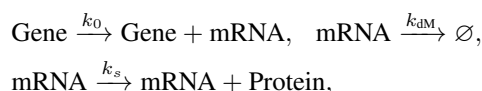
In this paper we develop and efficiently implement in iNA, the Inverse Omega Square (IOS) method which gives accurate estimates for the mean concentrations and variance / covariance of noise about them for systems whose molecule numbers vary over wide ranges (few to thousands of molecules). The new method is tested on two gene expression models involving bimolecular reactions. Remarkably, the results of a few seconds long IOS calculation is found to agree very well with those from hour long ensemble averaging of SSA trajectories.

II. RESULTS

In this section we describe the results of the novel IOS method implemented inside iNA, compare with the results of the RE, LNA and EMRE approximations of the CME and with exact stochastic simulations using the SSA and finally discuss the implementation and its performance. The three methods (LNA, EMRE, IOS) are derived from the system-size expansion (SSE) of the CME [2] which is applicable to monostable chemical systems. Technical details of the various approximation methods are provided in the section Methods.

A. Applications

We consider a two-stage model of gene expression with an enzymatic degradation mechanism:



Parameter	(i)	(ii)	(iii)
$k_0[G]$	$2.4\text{min}^{-1}\mu M$	$0.024\text{min}^{-1}\mu M$	$0.024\text{min}^{-1}\mu M$
k_{dM}	20min^{-1}	0.2min^{-1}	0.2min^{-1}
k_s	1.5min^{-1}	1.5min^{-1}	15min^{-1}
k_{-1}, k_2	2min^{-1}	2min^{-1}	20min^{-1}
k_1	$400(\mu M\text{min})^{-1}$	$400(\mu M\text{min})^{-1}$	$4000(\mu M\text{min})^{-1}$

Table I: Kinetic parameters used for the gene expression model, scheme 1, as discussed in the main text. The volume is fixed to $\Omega = 10^{-15}l$ with an enzyme concentration of $0.1\mu M$ corresponding to 60 enzyme molecules. The Michaelis-Menten constant is $0.01\mu M$ in all cases.



The scheme involves the transcription of mRNA, its translation to protein and subsequent degradation of both mRNA and protein. Note that while mRNA is degraded via an un-specific linear reaction, the degradation of protein occurs via an enzyme catalyzed reaction. The latter may model proteolysis, the consumption of protein by a metabolic pathway or other post-translational modifications. A simplified version of this model is one in which the protein degradation is replaced by the linear reaction: $\text{Protein} \rightarrow \emptyset$. Over the past decade the latter model has been the subject of numerous studies, principally because it can be solved exactly since the scheme is composed of purely first-order reactions [5]–[7]. However, the former model as given by scheme (1), cannot be solved exactly because of the bimolecular association reaction between enzyme and protein. Hence in what follows we demonstrate the power of approximation methods to infer useful information regarding the mechanism’s intrinsic noise properties.

We consider the model with three different parameter sets (see Tab I) and fix the compartment volume to one femtoliter (one micron cubed). For all three cases, the REs predict the same steady state mRNA and protein concentrations: $[\text{mRNA}] = 0.12\mu M$ and $[\text{Protein}] = 0.09\mu M$. These correspond to 72 and 54 molecule numbers, respectively.

We have used iNA to compute the mean concentrations using the REs and the variance of fluctuations using the LNA for parameter set (i) in which transcription is fast. Comparing these with SSA estimates (see Fig. 1) we see that the REs and LNA provide reasonably accurate results for this parameter set. This analysis was within the scope of the previous version of iNA [4]. However, the scenario considered is not particularly realistic. This is since the ratio of protein and mRNA lifetimes in this example is approximately 100 (as estimated from the time taken for the concentrations to reach 90% of their steady state values) while an evaluation of 1,962 genes in budding yeast showed that the ratios have median and mode close to 3 [7].

We now consider parameter set (ii), the case of moderate transcription. In Fig. 2(a) and (b) we compare the RE and LNA predictions of mean concentrations and variance of

fluctuations with that obtained from the SSA. Notice that in this case, the two are in severe disagreement. The SSA predicts that the mean concentration of protein is larger than that of the mRNA while the REs predict the opposite. It is also the case that the variance estimate of the LNA is considerably smaller than that of the SSA. In Fig. 2(c) we show the mean concentrations computed using the EMRE and the variance computed using the IOS method. Note that the latter are in good agreement with the SSA in Fig. 2(b). Note further that while the EMRE was already implemented in the previous version of iNA, the IOS was not. Hence the present version is the first to provide estimates of the mean concentrations and of the size of the noise which are more accurate than both the REs and the LNA. The transient times for this case are given by 37 minutes for protein and 12 minutes for mRNA concentrations with a ratio of approximately 3 in agreement with the median and mode of experimentally measured ratios. Hence this example provides clear evidence of the need to go beyond the RE and LNA level of approximation for physiologically relevant parameters of the gene expression model.

Finally we consider parameter set (iii), the case of moderate transcription and fast translation (k_s is an order of magnitude larger in (iii) compared to (ii)). Previous studies have shown that increased translation efficiency leads to increased noise in the protein abundance [5] due to the proteins being produced in bursts [7]. Indeed a single realization of the SSA shows proteins expressed in sharply peaked bursts (Fig. 3(a)). We used iNA to compute the mean concentrations of mRNA and protein according to RE, EMRE and IOS approximation methods (Fig. 3(b)). These are contrasted with the same mean concentrations computed from SSA simulations (Fig. 3(c)). Note that IOS provides the most accurate result, followed by the EMRE and the REs. The latter performs poorly because it does not take into account the effect of large fluctuations due to bursty behavior on the mean concentrations. Transient times of 48 minutes and 12 minutes for protein and mRNA concentrations have been extracted from the time course data as for previous cases; these are similar to those observed in the expression of the *E. coli* proteome [8], once again showing the necessity of approximation methods beyond the RE and LNA to study physiologically relevant cases.

B. Implementation

iNA’s framework consist of three layers of abstraction: the SBML parser which sets up a mathematical representation of the reaction network, a module which performs the SSE analytically using the computer algebra system Ginac [9] and a just-in-time (JIT) compiler based on the LLVM [10] framework which compiles the mathematical model into platform specific machine code at runtime yielding high performance of SSE and SSA analyses implemented in iNA.

As elaborated in the Methods section, in addition to the

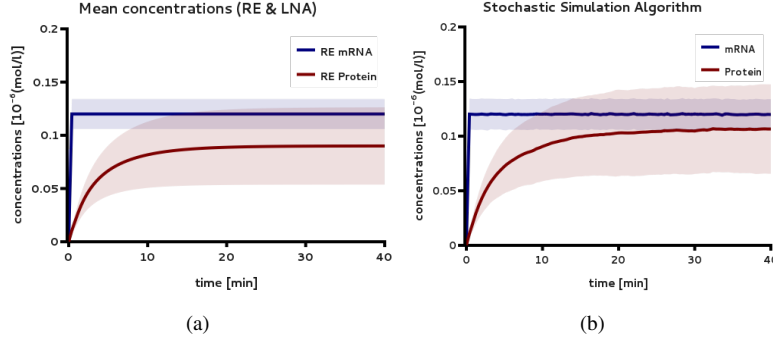


Figure 1: Gene expression model with fast transcription rate. In panels (a) and (b) we compare the RE predictions of mean concentrations with those obtained from ensemble averaging 3,000 SSA trajectories. The shaded areas denote the region of one standard deviation around the average concentrations which in (a) has been computed using the LNA and in (b) using the SSA. The results in (a) and (b) are in approximate agreement.

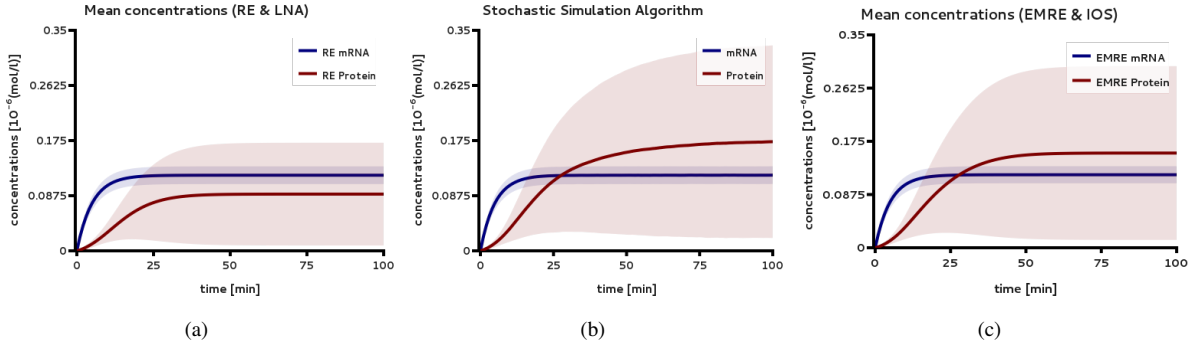


Figure 2: Gene expression model with moderate transcription rate. In panels (a) and (b) we compare the RE and LNA predictions of mean concentrations and variance of fluctuations with those obtained from ensemble averaging 30,000 stochastic realizations computed using the SSA. Note that the RE and LNA predictions are very different than the actual values. In panel (c) we show the mean concentration prediction according to the EMRE and the variance prediction according to IOS. These are in good agreement with those obtained from the SSA.

REs, van Kampen's SSE upon which the LNA, EMRE and IOS methods are based, requires the numerical solution of the following high dimensional system of linear equations

$$\frac{\partial \vec{x}_{\text{SSE}}}{\partial t} = \underline{B} \vec{x}_{\text{SSE}} + \vec{A}, \quad (2)$$

whose coefficients are automatically computed by iNA using the system size expansion from an SBML file. Note that the coefficient matrix $\underline{B} \equiv \underline{B}([\vec{X}])$ and the vector $\vec{A} \equiv \vec{A}([\vec{X}])$ depend parametrically on the solution of the REs. The vector \vec{x}_{SSE} is defined as $(\Omega^{-1} \vec{x}_{\text{LNA,EMRE}}, \Omega^{-2} \vec{x}_{\text{IOS}})$ where $\vec{x}_{\text{LNA,EMRE}}$ defines the variance/covariance of the LNA together with the corrections to mean concentrations of the REs according to the EMRE and \vec{x}_{IOS} defines the corrections to the variance/covariance of the LNA and the corrections to the mean concentrations of the EMRE according to the IOS. A particular simple result can be inferred by setting the time derivative of the REs and Eq. (2) to zero yielding $\underline{B} \vec{x}_{\text{SSE}}^* =$

$-\vec{A}$ whose solution is iNA's steady state analysis. First the typically nonlinear REs are solved using the Newton-Raphson method with line search [11] and consequently the solution to the latter linear equation is found using standard linear algebra. In the present version of iNA, time-course analysis is efficiently performed by means of the all-round ODE integrator LSODA which automatically switches between explicit and implicit methods [12].

The number of simultaneous equations to be solved for the LNA, EMRE analysis is approximately proportional to N^2 where N is the number of independent species after conservation analysis [13] while for IOS, the number of equations is approximately proportional to N^3 . The quadratic and cubic dependencies represent a challenge for software development. iNA's previous version addressed this need for performance by providing automated OpenMP parallelism and a bytecode interpreter (BCI) for efficient

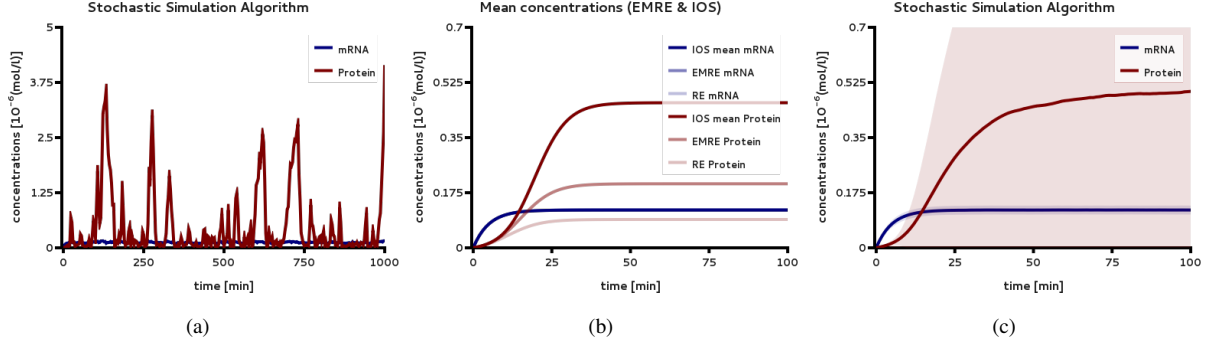
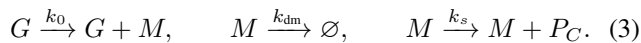


Figure 3: Gene expression model with efficient translation. In panel (a) we show a single SSA trajectory which illustrates the large protein bursts in the protein concentrations. In panel (b) we show the concentrations predicted by REs, EMRE and IOS methods and in (c) we show the mean concentrations obtained from ensemble averaging 30,000 SSA trajectories. Only the IOS method is in good quantitative agreement with the SSA predictions. These predictions are over 5 times larger than those of REs, the discrepancies stemming from the fact that the latter do not take into account contributions to the mean due to large fluctuations.

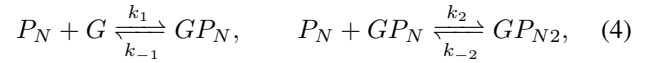
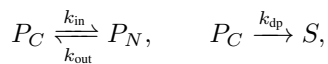
expression evaluation [4]. The latter concept has proven its performance for both SSE and SSA methods while maintaining compatibility over many platforms. With the present version we introduce a cross platform strategy for JIT compilation of the SSE or SSA methods. iNA uses the modern compiler framework LLVM providing fast JIT compilation that allows to emit and execute platform specific machine code at runtime [10]. This allows us to automatically compile the system size expansion ODEs as native machine code executable directly on the CPU without resorting to interpreters. Therefore iNA's JIT feature enjoys the speed of statically compiled code while maintaining the flexibility common to interpreters. Moreover, the LLVM compiler framework offers the use of optimizations on platform independent and platform specific instruction levels which become advantageous for computationally expensive calculations.

C. Performance

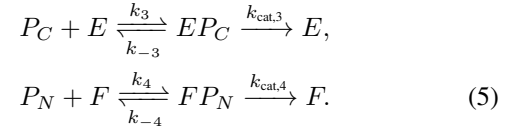
In order to benchmark the present implementation we consider an ensemble of independent single cell oscillators with weak negative feedback involving 10 species and 16 reactions. The model describes the transcription of mRNA, M , by a clock gene, G and its translation to a cytosolic clock protein, P_C :



A negative feedback loop arises from nuclear import of cytosolic protein and its binding to the gene which represses transcription:



where P_N denotes the nuclear protein. Note that protein P_C is consumed also by a different pathway involving S which is not explicitly modeled. As in the preceding example, degradation occurs via enzyme-catalyzed mechanisms, in this case via cytosolic and nuclear proteases E and F :



The model has been analyzed in detail in Ref. [4] using the EMREs computed by iNA. Therein it has been shown that intrinsic noise can amplify transient oscillations which are visible even at the cell population level. In Fig. 4 we compare the outcome of 50,000 independent realizations of the SSA with iNA's prediction of these synchronous noise-induced oscillations using the EMRE for the mean concentrations together with the IOS estimate of variance (the new feature in the present version of iNA). These estimates agree well with those obtained from simulations.

After conservation analysis the model comprises 7 species yielding a total number of 161 simultaneous equations for the SSE and hence is well suited for direct benchmarking purposes. This is particularly challenging for ODE integrators since the underlying stochastic dynamics causes the full system of 161 coupled equations to exhibit damped oscillations. The results of the benchmarks are summarized in Tab. II highlighting the performance of the present version of iNA over the previous version. The improvements of iNA's SSE using the LSODA over the previous Rosenbrock method reduces the execution time by a factor 2 – 3 while factors of 5 – 6 are attained using the JIT compiler which

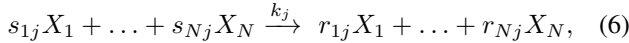
combined reduce the execution time from 25s to less than 2s. This is compared to the execution time of the SSA which is computationally extremely expensive because of the considerable number of trajectories which need to be averaged to obtain accurate statistics.

method	SSE, LSODA	SSE, Rosenbrock	SSA, single/ens.
BCI	12.88s (13.13s)	25.40s (25.66s)	0.15s/2.0h
BCI (opt.)	8.21s (8.51s)	26.59s (26.89s)	0.13s/1.8h
JIT	1.46s (1.86s)	10.77s (15.51s)	0.10s/1.4h
JIT (opt.)	1.12s (6.34s)	10.91s (17.04s)	0.10s/1.4h

Table II: Execution times of iNA for SSE and SSA for the gene oscillatory network reproducing Fig. 4 (a) and (b), respectively. The table compares the SSE method using the ODE integrator LSODA and Rosenbrock in combination with a bytecode interpreter (BCI) or iNA's JIT feature. The value in the brackets denotes the overall time including bytecode assembly or JIT compilation. The table also includes the execution times using optimizations (opt.) which speeds up execution times at the expense of longer compilation times. The SSA value denotes the average time for a single run while the values in the brackets denote the extrapolated value for an ensemble of 50,000 independent realizations needed to generate Fig. 4(b). Benchmarks were performed on MacOS 10.7, Core 2 Duo @1.4Ghz (64Bit) using one core.

III. METHODS

We consider a general reaction network confined in a volume Ω under well-mixed conditions and involving the interaction of N distinct chemical species via R chemical reactions of the type



where j is the reaction index running from 1 to R , X_i denotes chemical species i , k_j is the reaction rate of the j^{th} reaction and s_{ij} and r_{ij} are the stoichiometric coefficients. Note that our general formulation does not require all reactions to be necessarily elementary.

The CME gives the time-evolution equation for the probability $P(\vec{n}, t)$ that the system is in a particular mesoscopic state $\vec{n} = (n_1, \dots, n_N)^T$ where n_i is the number of molecules of the i^{th} species. It is given by:

$$\frac{\partial P(\vec{n}, t)}{\partial t} = \sum_{j=1}^R \left(\prod_{i=1}^N E_i^{-S_{ij}} - 1 \right) \hat{a}_j(\vec{n}, \Omega) P(\vec{n}, t), \quad (7)$$

where $S_{ij} = r_{ij} - s_{ij}$, $\hat{a}_j(\vec{n}, \Omega)$ is the propensity function such that the probability for the j^{th} reaction to occur in the time interval $[t, t + dt)$ is given by $\hat{a}_j(\vec{n}, \Omega) dt$ [3] and $E_i^{-S_{ij}}$ is the step operator defined by its action on a general function of molecular populations as $E_i^{-S_{ij}} g(n_1, \dots, n_i, \dots, n_N) = g(n_1, \dots, n_i - S_{ij}, \dots, n_N)$ [2].

The CME is typically intractable for computational purposes because of the inherent large state space. iNA circumvents this problem by approximating the moments of probability density solution of the CME systematically using van Kampen's SSE [2], [14]. The starting point of the analysis is van Kampen's ansatz

$$\frac{\vec{n}}{\Omega} = [\vec{X}] + \Omega^{-1/2} \vec{\epsilon}, \quad (8)$$

by which one separates the instantaneous concentration into a deterministic part given by the solution $[\vec{X}]$ of the macroscopic REs for the reaction scheme (6) and the fluctuations around it parametrized by $\vec{\epsilon}$. fluctuations around it. Note that $\vec{f} = \lim_{\Omega \rightarrow \infty} \vec{a}/\Omega$. The change of variable causes the probability distribution of molecular populations $P(\vec{n}, t)$ to be transformed into a new probability distribution of fluctuations $\Pi(\vec{\epsilon}, t)$. The time derivative, the step operator and the propensity functions are also transformed (see for [4] for details). In particular the propensities expand in powers of $\Omega^{-1/2}$ and are given by

$$\begin{aligned} \frac{\hat{a}_j(\vec{n}, \Omega)}{\Omega} &= f_j^{(0)}([\vec{X}]) + \Omega^{-1/2} \epsilon_\alpha \frac{\partial f_j^{(0)}([\vec{X}])}{\partial [X_\alpha]} + \Omega^{-1} f_j^{(1)}([\vec{X}]) \\ &\quad + \frac{1}{2} \Omega^{-1} \epsilon_\alpha \epsilon_\beta \frac{\partial f_j^{(0)}([\vec{X}])}{\partial [X_\alpha] \partial [X_\beta]} + O(\Omega^{-3/2}). \end{aligned} \quad (9)$$

Note that here we have used the convention that twice repeated Greek indices are summed over 1 to N , which we use in what follows as well. Consequently the CME up to order Ω^{-1} can be written as

$$\begin{aligned} \frac{\partial \Pi(\vec{\epsilon}, t)}{\partial t} &- \Omega^{1/2} \left(\frac{\partial [X_\alpha]}{\partial t} - \sum_{k=1}^R S_{\alpha k} f_k^{(0)}([\vec{X}]) \right) \partial_\alpha \Pi(\vec{\epsilon}, t) \\ &= \left(\Omega^0 \mathcal{L}^{(0)} + \Omega^{-1/2} \mathcal{L}^{(1)} + \Omega^{-1} \mathcal{L}^{(2)} \right) \Pi(\vec{\epsilon}, t) + O(\Omega^{-3/2}), \end{aligned} \quad (10)$$

where the operators are

$$\mathcal{L}^{(0)} = -\partial_\alpha J_\alpha^\beta \epsilon_\beta + \frac{1}{2} \partial_\alpha \partial_\beta D_{\alpha\beta}, \quad (11)$$

$$\begin{aligned} \mathcal{L}^{(1)} &= -\partial_\alpha D_\alpha^{(1)} - \frac{1}{2!} \partial_\alpha J_\alpha^{\beta\gamma} \epsilon_\beta \epsilon_\gamma + \frac{1}{2!} \partial_\alpha \partial_\beta J_{\alpha\beta}^\gamma \epsilon_\gamma \\ &\quad - \frac{1}{3!} \partial_\alpha \partial_\beta \partial_\gamma D_{\alpha\beta\gamma}, \end{aligned} \quad (12)$$

$$\begin{aligned} \mathcal{L}^{(2)} &= -\partial_\alpha J_\alpha^{(1)\beta} \epsilon_\beta + \frac{1}{2!} \partial_\alpha \partial_\beta D_{\alpha\beta}^{(1)} - \frac{1}{3!} \partial_\alpha J_\alpha^{\beta\gamma\delta} \epsilon_\beta \epsilon_\gamma \epsilon_\delta \\ &\quad + \frac{1}{2!} \frac{1}{2!} \partial_\alpha \partial_\beta J_{\alpha\beta}^{\gamma\delta} \epsilon_\gamma \epsilon_\delta - \frac{1}{3!} \partial_\alpha \partial_\beta \partial_\gamma J_{\alpha\beta\gamma}^\delta \epsilon_\delta \\ &\quad + \frac{1}{4!} \partial_\alpha \partial_\beta \partial_\gamma \partial_\delta D_{\alpha\beta\gamma\delta}, \end{aligned} \quad (13)$$

and the SSE coefficients are given by

$$\begin{aligned} D^{(n)}_{ij..r} &= \sum_{k=1}^R S_{ik} S_{jk} \dots S_{rk} f_k^{(n)}([\vec{X}]), \\ J^{(n)}_{ij..r}{}^{st..z} &= \frac{\partial}{\partial [X_s]} \frac{\partial}{\partial [X_t]} \dots \frac{\partial}{\partial [X_z]} D^{(n)}_{ij..r}. \end{aligned} \quad (14)$$

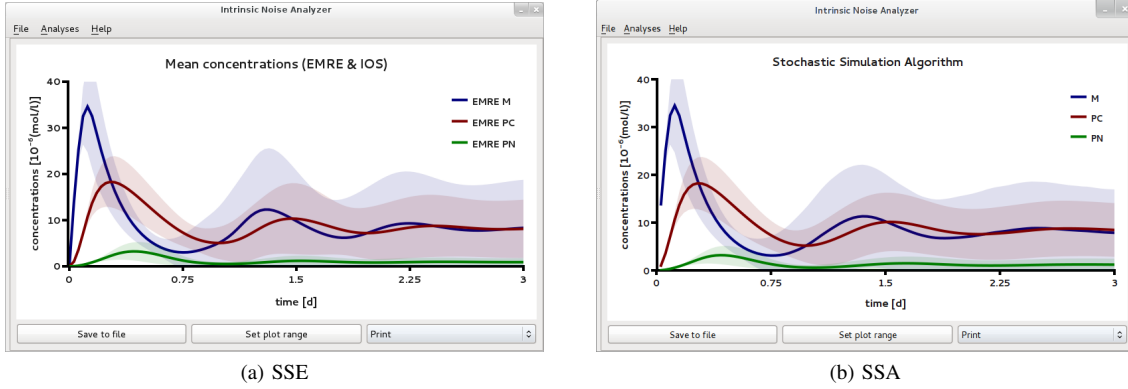


Figure 4: Screenshots of iNA's SSE using the EMRE and IOS analysis versus the average of 50,000 stochastic simulations of the gene oscillatory network which exhibits amplified synchronous oscillations which are not captured by the REs (see Fig. 8 in Ref. [4]). The EMRE mean concentrations together with the IOS variances in panel (a) are in good agreement with SSA simulations in panel (b). The computation of (a) takes less than two seconds compared to more than one hour for the stochastic simulation in (b), see Tab. II, which highlights the efficiency of the system size method implemented in iNA.

Note that the above expressions generalize the expansion carried out in Ref. [15] to include also nonelementary reactions as for instance trimolecular reactions or reactions with propensities of the Michaelis-Menten type [16]. Note also that the $\Omega^{1/2}$ term vanishes since the macroscopic REs are given by $\partial_t[X_\alpha] = \sum_{k=1}^R S_{\alpha k} f_k^{(0)}([\vec{X}])$ leaving us with a series expansion of the CME in powers of the inverse square root of the volume. In what follows we shall omit the upper index in the bracket of the SSE coefficients in the case of $n = 0$.

Linear Noise Approximation: We now proceed by constructing equations for the moments of the $\vec{\epsilon}$ variables. We follow the derivation presented in Ref. [15] and expand the probability distribution of fluctuations $\Pi(\vec{\epsilon}, t)$ in terms of the inverse square root of the volume

$$\Pi(\vec{\epsilon}, t) = \sum_{j=0}^{\infty} \Pi_j(\vec{\epsilon}, t) \Omega^{-j/2}. \quad (15)$$

As a consequence there exists an equivalent expansion of the moments

$$\langle \epsilon_k \epsilon_l \dots \epsilon_m \rangle = \sum_{j=0}^{\infty} [\epsilon_k \epsilon_l \dots \epsilon_m]_j \Omega^{-j/2}, \quad (16)$$

where the following definition has been used

$$[\epsilon_k \epsilon_l \dots \epsilon_m]_j = \int \epsilon_k \epsilon_l \dots \epsilon_m \Pi_j(\vec{\epsilon}, t) d\vec{\epsilon}. \quad (17)$$

Using the expansion of the probability density, Eq. (15), together with Eq. (10) and (11), we find after equating all terms of order Ω^0

$$\frac{\partial}{\partial t} \Pi_0(\vec{\epsilon}, t) = \mathcal{L}^{(0)} \Pi_0(\vec{\epsilon}, t) \quad (18)$$

which is a Fokker-Planck equation with linear drift and diffusion coefficients, also called the Linear Noise Approximation. If the initial state is deterministic, i.e., $\vec{n}/\Omega = [\vec{X}]$ initially, the solution is a multivariate Gaussian distribution [2] centered around $[\epsilon]_0 = 0$ for all times. The time evolution equations of the second moment are obtained by multiplying the latter by ϵ_i and $\epsilon_i \epsilon_j$ respectively and performing the integration over $\vec{\epsilon}$:

$$\frac{\partial}{\partial t} [\epsilon_i \epsilon_j]_0 = J_i^\alpha [\epsilon_\alpha \epsilon_j]_0 + \frac{1}{2} D_{ij} + (i \leftrightarrow j), \quad (19)$$

where $(i \leftrightarrow j)$ denotes the permutations of indices. It then follows that within the LNA all higher even moments can be expressed in terms of the second moment by virtue of Wick's theorem [17]:

$$[\epsilon_1 \epsilon_2 \dots \epsilon_\zeta]_0 = \sum_{\substack{\text{all possible pairings} \\ P \text{ of } \{1, 2, \dots, \zeta\}}} [\epsilon_{P_1} \epsilon_{P_2}]_0 \dots [\epsilon_{P_{\zeta-1}} \epsilon_{P_\zeta}]_0. \quad (20)$$

Note that all odd moments are zero since $\Pi_0(\vec{\epsilon}, t)$ is centered.

EMRE and IOS approximations: The system size expansion can be used to calculate higher order corrections to the moments. The leading order correction to the first moment is given by

$$\partial_t [\epsilon_i]_1 = J_i^\alpha [\epsilon_\alpha]_1 + \frac{1}{2} J_i^{\alpha\beta} [\epsilon_\alpha \epsilon_\beta]_0 + D_i^{(1)}, \quad (21)$$

which is equivalent to the EMRE [18] implemented in the previous version of iNA [4]. The correction to the second moment has been derived in detail in Ref. [15] and is given

by

$$\begin{aligned}\partial_t[\epsilon_i \epsilon_j]_2 &= J_i^\alpha[\epsilon_\alpha \epsilon_j]_2 + \frac{1}{2} J_i^{\alpha\beta}[\epsilon_\alpha \epsilon_\beta \epsilon_j]_1 + \frac{1}{3!} J_i^{\alpha\beta\gamma}[\epsilon_\alpha \epsilon_\beta \epsilon_\gamma \epsilon_j]_0 \\ &\quad + D_i^{(1)}[\epsilon_j]_1 + J_i^{(1)\alpha}[\epsilon_\alpha \epsilon_j]_0 + \frac{1}{4} J_{ij}^{\alpha\beta}[\epsilon_\alpha \epsilon_\beta]_0 \\ &\quad + \frac{1}{2} D_{ij}^{(1)} + (i \leftrightarrow j),\end{aligned}\quad (22)$$

which is inherently coupled to the leading order non-Gaussian correction of the third moment

$$\begin{aligned}\partial_t[\epsilon_i \epsilon_j \epsilon_k]_1 &= J_i^\alpha[\epsilon_\alpha \epsilon_j \epsilon_k]_1 + \frac{1}{2} J_i^{\alpha\beta}[\epsilon_\alpha \epsilon_\beta \epsilon_j \epsilon_k]_0 + D_i^{(1)}[\epsilon_j \epsilon_k]_0 \\ &\quad + \frac{1}{2} D_{jk}[\epsilon_i]_1 + \frac{1}{2} J_{jk}^\alpha[\epsilon_\alpha \epsilon_i]_0 + \frac{1}{3} D_{ijk} \\ &\quad + (i \leftrightarrow j) + (j \leftrightarrow k).\end{aligned}\quad (23)$$

Note that the above equations depend on the fourth moment which is readily evaluated using Wick's theorem:

$$[\epsilon_i \epsilon_j \epsilon_k \epsilon_l]_0 = [\epsilon_i \epsilon_j]_0 [\epsilon_k \epsilon_l]_0 + [\epsilon_i \epsilon_k]_0 [\epsilon_j \epsilon_l]_0 + [\epsilon_i \epsilon_l]_0 [\epsilon_j \epsilon_k]_0. \quad (24)$$

In order to relate the above moments back to the moments of the concentration variables we use Eqs. (8) and (16) to find expressions for the mean concentrations and covariance of fluctuations which are given by

$$\begin{aligned}\left\langle \frac{n_i}{\Omega} \right\rangle &= [X_i] + \Omega^{-1}[\epsilon_i]_1 + O(\Omega^{-2}) \\ \Sigma_{ij} &= \left\langle \left(\frac{n_i}{\Omega} - \left\langle \frac{n_i}{\Omega} \right\rangle \right) \left(\frac{n_j}{\Omega} - \left\langle \frac{n_j}{\Omega} \right\rangle \right) \right\rangle \\ &= \Omega^{-1}[\epsilon_i \epsilon_j]_0 + \Omega^{-2}([\epsilon_i \epsilon_j]_2 - [\epsilon_i]_1 [\epsilon_j]_1) + O(\Omega^{-3})\end{aligned}\quad (25)$$

The Ω^{-1} term of the latter equation gives the LNA estimate for the covariance while including terms to order Ω^{-2} gives the IOS (Inverse Omega Squared) estimate.

Note that by inspection of Eqs. (21) and (22) it follows these corrections are nonzero if the reaction network involves at least one bimolecular reaction. Moreover we can also estimate the next order correction to the mean corrections, i.e., the correction to the EMRE, as has been done in Ref. [15]

$$\begin{aligned}\partial_t[\epsilon_i]_3 &= J_i^\alpha[\epsilon_\alpha]_3 + \frac{1}{2} J_i^{\alpha\beta}[\epsilon_\alpha \epsilon_\beta]_1 \\ &\quad + \frac{1}{3!} J_i^{\alpha\beta\gamma}[\epsilon_\alpha \epsilon_\beta \epsilon_\gamma]_1 + J_i^{(1)\alpha}[\epsilon_\alpha]_1.\end{aligned}\quad (27)$$

The solution of the above equation allows us to obtain the mean concentration accurate to order Ω^{-2}

$$\left\langle \frac{n_i}{\Omega} \right\rangle = [X_i] + \Omega^{-1}[\epsilon_i]_1 + \Omega^{-2}[\epsilon_i]_2 + O(\Omega^{-3}), \quad (28)$$

where the first term is the RE solution, the first two terms constitute the EMRE estimate and including all three terms gives the IOS estimate for the mean concentrations.

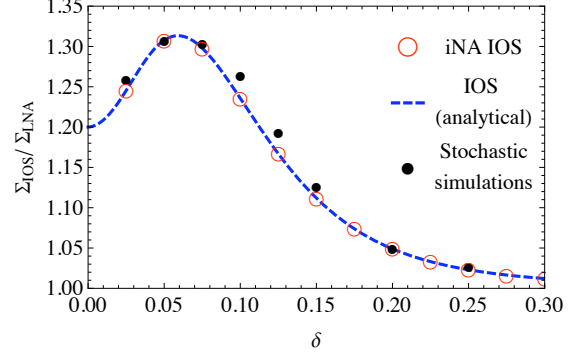


Figure 5: Michaelis-Menten reaction. We have verified the soundness of our numerical implementation by comparing with the analytical result using the IOS derived in Ref. [19]. The graph shows the ratio of IOS and LNA variance of substrate fluctuations against the ratio δ of the free enzyme concentration and the total enzyme concentration in steady state conditions. This is also compared to the SSA where the ratio of SSA and LNA variance has been used.

Summarizing the system size expansion is obtained by defining the vector $\vec{x}_{SSE} = (\Omega^{-1}\vec{x}_{LNA,EMRE}, \Omega^{-2}\vec{x}_{IOS})$ with components

$$\begin{aligned}\vec{x}_{LNA,EMRE} &= \text{vec}([\epsilon_i \epsilon_j]_0, [\epsilon_i]_1), \\ \vec{x}_{IOS} &= \text{vec}([\epsilon_i \epsilon_j \epsilon_k \epsilon_l]_0, [\epsilon_i \epsilon_j \epsilon_k]_1, [\epsilon_i \epsilon_j]_2, [\epsilon_i]_3),\end{aligned}\quad (29)$$

where the symbol $\text{vec}(\cdot)$ denotes the vectorization with respect to the independent components of the involved fully symmetric tensors. Then the vector $\vec{x}_{LNA,EMRE}$ defines the LNA covariance together with the EMRE and \vec{x}_{IOS} defines the corrections to the LNA covariance and third moments of the SSE together with the order Ω^{-2} correction to the EMRE mean concentrations. It then follows that \vec{x}_{SSE} satisfies the linear ODE

$$\frac{\partial \vec{x}_{SSE}}{\partial t} = \underline{B} \vec{x}_{SSE} + \vec{A}, \quad (30)$$

where the upper block triangular matrix $\underline{B} \equiv \underline{B}([\vec{X}])$ and the full vector $\vec{A} \equiv \vec{A}([\vec{X}])$ are determined by Eqs. (19,21,22,23,27) and depend parametrically on the solution of the REs.

In Fig. 5 we have verified the correctness of iNA's numerical implementation of the IOS approximation by comparing with the compact IOS analytical expression recently derived for the case of a simple enzyme catalyzed reaction (see Eq. (74) in Ref. [19]).

IV. DISCUSSION

In this article we have introduced and implemented the IOS approximation in the software package iNA. This allows the mean concentrations and variances to be determined accurate to order Ω^{-2} , an approximation which is superior

to the previously implemented methods of LNA (variances accurate to order Ω^0) and EMRE (mean concentrations accurate to order Ω^{-1}). As we shown this increased accuracy is desirable to accurately account for the effects of intrinsic noise in biochemical reaction networks under low molecule number conditions. In particular, we have demonstrated the utility of the software by analyzing an example of gene expression with a functional enzyme and showcased the efficiency of the method using a more complex gene oscillatory network. We have also extended iNA by a more efficient JIT compilation strategy in combination with improved numerical algorithms which offers high performance and enables computations feasible even on desktop PCs. This feature is particularly important when analyzing noise in reaction networks of intermediate or large size with well-separated timescales and some bimolecular reactions, conditions that have been shown to amplify the deviations from the conventional rate equation description [20].

We conclude by pointing out that while stiffness arising from multiple timescales in biological networks is still an insufficiently resolved problem of stochastic simulation methods [3], [21], it does not pose significant problems for the LNA, EMRE and IOS approximation methods. This is since the implementations of the latter in iNA are able to deal with stiffness natively using well established implicit methods developed for ordinary differential equations.

ACKNOWLEDGMENT

RG gratefully acknowledges support by SULSA (Scottish Universities Life Science Alliance).

AVAILABILITY

The software iNA version 0.3 is freely available under <http://code.google.com/p/intrinsic-noise-analyzer/> as executable binaries for Linux, MacOSX and Microsoft Windows, as well as the full source code under an open source license.

REFERENCES

- [1] Y. Ishihama, T. Schmidt, J. Rappsilber, M. Mann, U. Hartl, M. J. Kerner, and D. Frishman, "Protein abundance profiling of the escherichia coli cytosol," *BMC Genomics*, vol. 9, p. 102, 2008.
- [2] N. van Kampen, *Stochastic processes in physics and chemistry*, 3rd ed. North-Holland, 2007.
- [3] D. Gillespie, "Stochastic simulation of chemical kinetics," *Annu. Rev. Phys. Chem.*, vol. 58, pp. 35–55, 2007.
- [4] P. Thomas, H. Matuschek, and R. Grima, "intrinsic noise analyzer: a software package for the exploration of stochastic biochemical kinetics using the system size expansion," *PlosOne*, p. (in press), 2012.
- [5] E. Ozbudak, M. Thattai, I. Kurtser, A. Grossman, and A. van Oudenaarden, "Regulation of noise in the expression of a single gene," *Nature genetics*, vol. 31, no. 1, pp. 69–73, 2002.
- [6] J. Paulsson, "Models of stochastic gene expression," *Physics of life reviews*, vol. 2, no. 2, pp. 157–175, 2005.
- [7] V. Shahrezaei and P. Swain, "Analytical distributions for stochastic gene expression," *Proceedings of the National Academy of Sciences*, vol. 105, no. 45, p. 17256, 2008.
- [8] Y. Taniguchi, P. Choi, G. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. Xie, "Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells," *Science*, vol. 329, no. 5991, pp. 533–538, 2010.
- [9] C. Bauer, A. Frink, and R. Kreckel, "Introduction to the ginac framework for symbolic computation within the c++ programming language," *Journal of Symbolic Computation*, vol. 33, no. 1, pp. 1–12, 2002.
- [10] C. Lattner and V. Adve, "Llvm: A compilation framework for lifelong program analysis & transformation," in *Code Generation and Optimization, 2004. CGO 2004. International Symposium on*. IEEE, 2004, pp. 75–86.
- [11] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical recipes: the art of scientific computing*, 3rd ed. Cambridge University Press, 2007.
- [12] L. Petzold, "Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations," *SIAM Journal on Scientific and Statistical Computing*, vol. 4, no. 1, pp. 136–148, 1983.
- [13] R. Vallabhajosyula, V. Chickarmane, and H. Sauro, "Conservation analysis of large biochemical networks," *Bioinformatics*, vol. 22, no. 3, pp. 346–353, 2006.
- [14] N. Van Kampen, "The expansion of the master equation," *Advances in Chemical Physics*, pp. 245–309, 1976.
- [15] R. Grima, P. Thomas, and A. Straube, "How accurate are the nonlinear chemical fokker-planck and chemical langevin equations?" *The Journal of Chemical Physics*, vol. 135, p. 084103, 2011.
- [16] C. Rao and A. Arkin, "Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the gillespie algorithm," *The Journal of chemical physics*, vol. 118, p. 4999, 2003.
- [17] J. Zinn-Justin, *Phase Transitions and Renormalization Group*. Oxford University Press, 2007.
- [18] R. Grima, "An effective rate equation approach to reaction kinetics in small volumes: Theory and application to biochemical reactions in nonequilibrium steady-state conditions," *The Journal of Chemical Physics*, vol. 133, p. 035101, 2010.
- [19] "A study of the accuracy of moment-closure approximations for stochastic chemical kinetics," *The Journal of Chemical Physics*, vol. 136, p. 154105, 2012.
- [20] P. Thomas, A. Straube, and R. Grima, "Stochastic theory of large-scale enzyme-reaction networks: Finite copy number corrections to rate equation models," *Journal of Chemical Physics*, vol. 133, no. 19, p. 195101, 2010.
- [21] Y. Cao, D. Gillespie, and L. Petzold, "Adaptive explicit-implicit tau-leaping method with automatic tau selection," *The Journal of Chemical Physics*, vol. 126, p. 224101, 2007.